Audible Outputs of Reading Machines for the Blind*

Franklin S. Cooper, Jane H. Gaitenby, Ignatius G. Mattingly,† Patrick W. Nye, and George N. Sholes
Haskins Laboratories, New Haven

The goal of research on reading machines for the blind at Haskins Laboratories is to produce by machine methods an output of clear, audible English from an input of ordinary printed text. The core problem—generating acceptable speech from phonetic spellings—seems very near a successful solution through synthesis-by-rule methods. There is still much to be done by way of evaluating and improving the synthetic speech, but the research can now turn to some of the other problems involved in setting up a complete Reading Service Center for the blind. During the six months covered by this report, attention has been focused on two main endeavors: evaluation studies of the reading machine output have continued with blind students, and further progress has been made toward automation of the entire print-to-speech generating process.

Evaluation by Blind Students

Continuing the work reported in the previous issue of the Bulletin, two studies have been made of student reactions to hearing some of their regular textbook assignments in the medium of synthetic speech. For the first study, with the help of faculty at the University of Connecticut, ten recorded passages totaling 2-1/2 hours of listening time were administered to six blind students. These passages covered chapters in psychology and psychiatry as well as ancient and modern literature. The content fell broadly into two classes: either basically simple prose style or more elaborate composition demanding close analytical attention.

Following these trial readings, the comments of the blind students showed general agreement on five points. First, all the students found the speech intelligible, and although an occasional word was missed, they had no trouble in following the meaning of the simple prose; however, some students found difficulty in concentrating on the subject matter of the more complex material. Second, all students were favorably impressed by the stress and intonation aspects of the speech. Third, all students complained about the "cold-in-the-head" quality of the speech, but the samples used were too short to determine whether the students would acclimate to this aspect of voice quality. Fourth, all students thought that the speed of presentation of the samples was too slow. [The rates ranged from 109 to 156 words per minute. The

---

† Also University of Connecticut, Storrs.
Latter is within the normal of human speaking rates but the long silences (2 to 8 sec.) between some sentences in these early recordings made the overall rate seem slow. These undue silences were eliminated in subsequent recordings. Finally, long and often unfamiliar polysyllabic words were recognized easily. The words missed were usually monosyllables embedded in sequences of other short words.

In a second study during the late fall of 1971, another series of tapes was prepared at substantially higher word rates (164-221 wpm). These tapes received the benefit of more recent refinements in the rules for producing and recording synthetic speech. The most detailed comments on the second tests were obtained from two female students who voiced opposing views that were, however, typical of the group as a whole. Both students noted the improvements in naturalness compared with the earlier tapes. The first student indicated that, for passages that were complex (in topic, grammar, or vocabulary), she might well have preferred a slower rate. This student noted that her difficulty in focusing attention on the content (rather than on the voice quality) might disappear with longer experience in listening, but she was uncertain about how well she could use synthetic speech as a primary study tool. The second student, who listened to a text having a simple narrative style, was enthusiastic. She claimed to have missed only two words in a 15-minute recording that "spoke" at 221 wpm. She felt that she could make use of synthetic speech as a primary study tool.

Plans for Further Testing and for a Reading Service Center

More sophisticated tests are scheduled for the spring of 1972. In addition, a faculty committee at the University of Connecticut is now actively planning further steps toward the development of a Reading Service Center which will be located on the campus and will utilize the Haskins Laboratories speech synthesis facilities. These plans call for a two-part program commencing with a 12- to 18-month study of the human, economic, and technical factors involved in the operation of such a Reading Service Center. Haskins Laboratories will be involved in this study as a supplier of synthetic speech material to the University, using the automated facility currently being developed. The University researchers will be responsible for distributing the tape recordings and conducting sequential listening tests, both with blind students (some of them veterans) enrolled at the University and with blind students in schools and colleges throughout Connecticut. The second part of the program will incorporate the data from these studies to make decisions on the size and type of computer and optical character recognition equipment required for an on-campus Reading Service Center and to seek funding for its implementation during the 1973-74 academic year.

Automating Text Preparation

At the Laboratories, the task of automating the production of synthetic speech from an input of printed text continues. Enquiries are in progress toward the acquisition (on lease) of a limited-font optical character recognizer. Needed is a suitable machine for converting text that has been typed in OCR-A or -B (upper and lower case) type face into an alphanumeric code on magnetic tape. (The choice of a "simple" OCR machine and human typists for the initial production phase is based primarily on cost considerations.)
Multifont machines to read book pages directly are available and their higher cost will be justified when a higher level of text production is wanted.) Optical character recognition represents the first of six stages involved in the production of a synthetic speech recording. These stages are shown in Figure 1.

Following the recognition stage, the words of the text are converted into phonemic form by means of a dictionary look-up (Stage 2). This dictionary now contains about 150,000 entries which are distributed in three compartments, with room for several-fold expansion. The first compartment contains a few hundred of the most frequently used words such as "the," "of," etc. In the second compartment is stored the overwhelming bulk of all entries. To facilitate access, this main store has been divided into functional "pages," which are referenced from a page-size table of contents. Locating an entry in the main store entails a two-part search, first through the table of contents, then through a page. The third compartment contains all oversize words (length greater than sixteen letters).

Each word entry, in both the high-frequency and main stores, contains the orthographic spelling, the phonetic respelling, and an indication of the word's usual grammatical functions. The initial version of the main store has now been completed, and programs for searching it are being written. These programs allow for editorial intervention to introduce new words that are not now available in the dictionary, as well as to correct errors.

**Stress and Intonation**

In the third stage, the phonemic string generated by the dictionary search is processed to introduce the stress and intonation features required to guide the synthesis program. Each dictionary word is (by the rules we are using) a member of one of five main stress classes: Low Stable, Low Unstable, Mid Unstable, High Stable, High Unstable. (Words with unstable stress shift their stress grade in specified contexts.) In general, low-stressed words are the so-called function words of speech (articles, prepositions, auxiliary verbs, many pronouns, connectives); words with mid stress are modifiers and verbs in the past tense (and past participles); high-stressed words are nouns (or multi-use words that can be nouns), words of four or more syllables, numerals, certain emphatic words, comparative and superlative forms of adjectives, and a small number of semantically special words that tend to receive full stress in normal speech.

In the fourth stage, the phonetic strings from Stage 2 and the stress and intonation assignments from Stage 3 are combined into a series of syllable-generating digital instructions by the computer program. These instructions are realized as a synthetic speech wave form by the synthesizer (Stage 5) which is recorded as a series of audible sentences in the final stage.

Recent work has centered on adjusting the specifications for the basic American English sounds (the phonemes) for better compatibility at fast word rates (above 150 wpm), modifying the speech program to provide pauses of various lengths, and refining the stress assignment rules for complex texts.
THE TEXT-TO-SPEECH PROCESSOR

1. OPTICAL CHARACTER READER

2. TEXT-TO-PHONEME DICTIONARY LOOK-UP

3. STRESS AND INTONATION ASSIGNMENT

4. SYNTHESIS-BY-RULE

5. PARALLEL RESONANCE SYNTHESIZER

6. AUDIO TAPE RECORDER

Fig. 1

Input typed in OCR-A typeface is read.

Computer store of 150,000 words is consulted.

Computer program punctuates text.

Program computes control signals.

Hardware device generates speech.

Speech is recorded on magnetic tape.
Duration is the prosodic feature of main interest currently. Problems involving the voice quality of the synthesizer, noted by the blind test subjects, are also under study.

Modifications of the Synthesis Program

The computer program incorporating the rules for synthesis has disc storage and retrieval facilities for text that has been entered via a phonetic keyboard. Recent modifications to this program have been of two types. First, powerful editing functions have been provided permitting text additions, deletions, and corrections to be made quickly and easily. While the operator performs these functions, the phonetic characters are displayed on a storage oscilloscope. In addition, the program has been expanded to allow fully automatic synthesis of long passages of printed material by arranging for the control of intersentence and interparagraph pauses.

In summary, substantial progress has been made toward the development of an automatic dictionary to transform printed material into phonetic equivalents, with stress and intonation provided by formula. Additional improvements have been made to the synthesis system and more are scheduled. Current research is directed toward 1) improving the naturalness of synthetic speech so that it will be comfortable to listen to when delivered at rates well above 150 wpm; and 2) producing synthetic speech directly from text automatically in the quantities needed for really adequate field trials.